STATISTICS IN MEDICINE
Statist. Med. 2005; **00**:1–10 *Prepared using* simauth.cls *[Version: 2002/09/18 v1.11]*

Bayesian sample size determination in non-sequential clinical trials: statistical aspects and some regulatory considerations

Jean-Marie Grouin[1], Maylis Coste[2], Pierre Bunouf[3], Bruno Lecoutre[4]

[1] *Université de Rouen and AFSSaPS,* [2] *I.R.I. Servier,* [3] *I.R.P. Fabre,* [4] *C.N.R.S. and Université de Rouen*

SUMMARY

The most common Bayesian methods for sample size determination are reviewed in the non sequential context of a confirmatory phase III trial in drug development. After recalling the regulatory viewpoint about sample size determination, we discuss the relevance of the various priors applied to the planning of clinical trials. We then investigate whether these Bayesian methods could compete with the usual frequentist approach to sample size determination and be considered as acceptable from a regulatory viewpoint. Copyright © 2005 John Wiley & Sons, Ltd.

The views expressed in this paper are not necessarily those of the AFSSaPS.

1. INTRODUCTION

In drug development, sample size determination (SSD) plays a crucial role, especially in the planning of phase III clinical trials, and sometimes in phase II trials when the range of effective and safe doses is to be identified at this stage.

Sample size determination is only referred to specifically in Section 3.5 of the ICH E9 guideline [1] in the context of a given trial. This guideline also refers to plans that should specify the ordered programme of clinical trials, 'with appropriate decision points and flexibility to allow modifications as knowledge accumulates'. Therefore it is acknowledged that overall and definitive preplanning is not always feasible nor desirable, and this may explain why the important issue of the required sample size is only addressed in the context of a trial.

The guideline focuses on the power-based Neyman-Pearson frequentist approach for determining the appropriate sample size to achieve the primary objective of the trial. Although it states that 'it is important to investigate the sensitivity of the sample size estimate to a variety of deviations from (the) assumptions', no mention is made of Bayesian methods.

This paper presents a critical review of the most common Bayesian methods for SSD, and focuses on the non sequential context of a confirmatory phase III trial in drug development.

In contrast to the frequentist approach which relies on the choice of a suitable value for the parameter of interest, and possibly for the parameter of nuisance, the Bayesian approach is able

*Correspondence to: jean-marie.grouin@univ-rouen.fr

to take into account the uncertainty of these parameters by generating their prior distribution from previously available information.

There are three main groups of Bayesian methods for SSD: (a) Predictive (or averaged) power methods [2-9] that consider the predictive probability of achieving the primary objective of a trial, given the available information. They can be seen as a hybrid of frequentist and Bayesian methods. (b) Interval length methods [10-13] that exclusively focus on the length of credibility intervals. (c) Decision-theoretic methods [14-21] that are based on the maximization of the expected utility. They are considered by their proponents as the only full Bayesian methods.

In Section 2 comment is made on the required sample size from the regulatory viewpoint. In Section 3, after recalling the Bayesian paradigm, we discuss the relevance of the various priors applied to the planning of clinical trials. In Section 4 we review the three groups of Bayesian methods and we investigate whether they could compete with the usual frequentist approach to SSD and be considered as acceptable from a regulatory viewpoint. The key ideas developed in this paper are summarised in Section 5 where the discussion focuses on the relevance and the practicality of the use of predictive probabilities for designing Phase III clinical trials in the context of drug development.

2. SOME GENERAL CONSIDERATIONS ON THE REQUIRED SAMPLE SIZE FROM THE REGULATORY PERSPECTIVE

From a regulatory view, confirmatory trials should be large enough to provide firm evidence of both efficacy and safety. Ideally, the optimal sample size of a confirmatory trial would be the minimum number of patients required to estimate the true treatment effect size on the primary endpoint with enough precision, while ensuring that there is enough data for the assessment of safety and possibly for the key-secondary efficacy endpoints.

In the context of a superiority trial and according to ICH E9 [1,p19], the necessary sample size should satisfy the power requirement for the primary objective of the trial, and for this purpose, the sponsor must specify beforehand a treatment effect size. The guideline recommends two alternatives for selecting this effect size: it 'may be based (either) on a judgment concerning the minimal effect which has clinical relevance... or on a judgment concerning the anticipated effect of the new treatment, where this is larger'.

It is worth noting that the point of view expressed in the guideline is not consensual. For example, Whitehead [22] insists that the effect chosen beforehand 'is in no way a guess at the true value' of the treatment effect. 'It is set, not by considering what value is believable, but by considering what value it is important to detect'.

The first alternative recommended by the guideline is to base the calculation of the sample size on the minimal clinically relevant effect Δ_{min}. This ensures an adequate power to detect an effect equal to, or larger than, this minimal effect, and this is always acceptable to the regulator. However, it is worth noting that even when the effect is statistically significant, if it is less than the minimal effect, it is not likely to obtain drug approval. This latter undesirable situation could be avoided if the hypotheses tested were $H_0 : \delta = \Delta_{min}$ versus $H_a : \delta > \Delta_{min}$ instead of $H_0 : \delta = 0$ versus $H_a : \delta > 0$. Nonetheless to demand the testing of these hypotheses for regulatory approval would require a larger sample size.

It is well known that most sponsors do not want to risk sizing their confirmatory phase III trials based on the minimal effect. The true effect is either larger than the minimal effect and

the sample size based on this latter is larger than it has to be, or the true effect is smaller and there is no need at all for a clinical trial. Therefore sponsors usually prefer the second alternative which is to base the calculation of the sample size 'on a judgment concerning the anticipated effect of the new treatment', hence on a guess at the true value.

This guess is educated when it relies on the results of earlier studies, but sometimes it is artificially derived in a reverse manner: the sample size is dictated by the resources available and the effect size to be detected is chosen to satisfy this sample size. In either case, if the guess is wrong, i.e. if the true effect is smaller than the anticipated effect, then the outcome of the trial is likely to be negative. This would clearly be a waste of time and resources for the sponsor and also unethical since patients may have been exposed to a possible risk in an undersized trial. Even in the case of a positive outcome, but where the observed effect is smaller than the anticipated effect, the regulatory authorities could suspect that the true effect is actually smaller than the anticipated effect and therefore could require this positive outcome to be replicated in a second adequately powered trial.

Therefore the need for a better educated guess at the true treatment effect still remains a relevant issue.

3. RELEVANT PRIORS FOR THE PLANNING OF CLINICAL TRIALS FROM THE REGULATORY PERSPECTIVE

Instead of considering only one specific value for the guess at the true treatment effect δ, the Bayesian approach considers a prior distribution on δ.

Suppose x denotes the data from a trial, then the Bayesian formula states that the posterior distribution $p(\delta|x)$ is proportional to the product of the likelihood $p(x|\delta)$ and the prior $p(\delta)$:

$$p(\delta|x) \propto p(x|\delta) \times p(\delta)$$

Inferences about δ are made from the posterior distribution, and since this is influenced by the prior, the choice of the prior is of crucial importance when reporting the results as well as planning the trial.

From a regulatory viewpoint, it is obvious that the reporting of the results, at least in the primary analysis planned in the protocol, should be as objective as possible, relying on the trial data only. Regarding the planning of the trial, the ICH E9 guideline [1] recommends that the basis of 'the estimates of any quantities used in the calculations (such as variances, ... the difference to be detected) should be given. In confirmatory trials, assumptions should normally be based on published data or on the results of earlier trials'. Consequently, subjective priors which reflect expert viewpoints (cf. Spiegelhalter et al. [3-6]) will probably not be accepted by the regulators and the prior distribution, which the regulators would agree on, is likely to be a posterior distribution derived only from relevant previous data. Therefore a prior distribution must also be chosen for these previous data and a reasonable solution is to select a non-informative prior (cf. Press [23] or Berger [24] for the definition of such reference priors). Another approach suggested [6] is to use a sceptical prior to harden the often too enthusiastic results from the previous trial, especially when this latter is a pilot phase II trial and not a confirmatory phase III trial. However, applying such sceptical priors to the SSD would require larger sample sizes than those based on non-informative priors, and consequently, sponsors might not be inclined to use them on a regular basis.

4. BAYESIAN APPROACHES TO THE PLANNING OF CLINICAL TRIALS

We review the three common Bayesian methods for SSD.

Predictive power methods

These are certainly the most known methods in the specific context of clinical trials [2-9]. The reason is that their criterion for determining the sample size is the usual frequentist power averaged over the possible values of the true effect according to its prior distribution. As the SSD involves in a straightforward manner the predictive probability of rejecting the null hypothesis from the trial data, the method is also known as the predictive power method.

In accordance with regulatory expectations, it is assumed that the rejection of the null hypothesis will be based on the trial data only, i.e. the frequentist analysis takes place when reporting the results, whereas the prior distribution based on previously available data is used for the planning only.

This approach can be illustrated in the context of a superiority trial where the treatment effect δ is expressed as the difference between two independent normal means with common unknown variance σ^2. Let us assume without any loss of generality that the sample size n is equal for both groups and denote d and s^2 as the estimated difference and common variance.

The test hypotheses considered are

$$H_0 : \delta = \delta_0 \text{ versus } H_a : \delta > \delta_0$$

where δ_0 can be chosen equal to 0 or Δ_{min}, as mentioned in Section 2.

Thus, n is determined such that the predictive probability (P_{pred}) of rejecting H_0 is at least greater than a given guarantee γ ($\gamma > 0.5$):

$$P_{pred}(\text{reject } H_0) = P_{pred}(t = \frac{d - \delta_0}{s\sqrt{\frac{2}{n}}} \geq t_{\nu,\alpha}) \geq \gamma$$

where $t_{\nu,\alpha}$ is the α upper point of a Student t distribution with $\nu = 2(n-1)$ degrees of freedom.

We are therefore interested in deriving the predictive density of the usual Student t statistic. Let us assume for (δ, σ) a standard conjugate prior distribution characterized by

$$\delta|\sigma^2 \sim N\left(d_0, 2\frac{\sigma^2}{n_0}\right) \text{ and } \sigma^2 \sim s_0^2 \left(\frac{\chi^2_{\nu_0}}{\nu_0}\right)^{-1}$$

This prior distribution can be considered as the posterior revised from the previous relevant data, assuming the usual non-informative prior. d_0, s_0^2, n_0, ν_0 respectively denote the estimated difference and common variance, the common group size and the number of degrees of freedom for these previous data.

It can easily be shown [7] that the predictive distribution of t given σ^2 is a non-central t' distribution.

$$t = \frac{d - \delta_0}{s\sqrt{\frac{2}{n}}}|\sigma^2 \sim \sqrt{\frac{n_0 + n}{n_0}} \times t'_\nu \left(\frac{d_0 - \delta_0}{\sigma\sqrt{2(\frac{1}{n_0} + \frac{1}{n})}}\right)$$

This distribution can easily be used to determine the sample size for a given variance (for example, $\sigma^2 = s_0^2$).

It should be clarified that considering an unknown variance can legitimately be seen as an unnecessary sophistication in the context of the SSD, since the solutions based on the known variance are generally very close, especially when the required sample sizes are high, which is usually the case in confirmatory phase III trials. However, since computations remain easily tractable, the assumption of a known variance can be relaxed without any inconvenience. The solution based on the unknown variance uses the marginal predictive density $p(t)$ of t :

$$p(t) = \int p(t|\sigma^2)p(\sigma^2)d\sigma^2$$

This involves a new distribution called K' by Lecoutre [7,8] who gave an explicit form for the density $p(t)$ and the distribution function. Grouin and Lecoutre [9] applied it to the context of the SSD.

Two particular cases are of practical importance.

$$\text{When } n \to \infty, \qquad P_{pred}(\text{reject } H_0) \to 1 - Pr(t_{\nu_0} > \frac{d_0 - \delta_0}{s_0\sqrt{\frac{2}{n_0}}}) = 1 - p_0$$

where t_{ν_0} is the usual t-distribution with $\nu_0 = 2(n_0 - 1)$ degrees of freedom, and p_0 is the one-sided level of significance based on previous data. For example, if $p_0 = 0.15$, this means that the predictive probability cannot exceed 85% whatever the sample size.

$$\text{When } n_0 \to \infty, \qquad \delta \to d_0 \text{ and } \sigma^2 \to s_0^2$$

i.e. δ and σ^2 are fixed as in the frequentist context and the predictive power is the usual power

$$P_{pred}(\text{reject } H_0) = Pr\left(t'_\nu(\frac{d_0 - \delta_0}{s_0\sqrt{\frac{2}{n}}}) > t_{\nu,\alpha}\right)$$

The sample size required with the predictive power approach is obviously much larger than the one obtained with the simple power approach. For example, suppose a pilot trial with $n_0 = 50$ patients per group where the estimated difference $d_0 = 2$ and the standard deviation $s_0 = 6$. The point of interest is to test $H_0 : \delta = 0$ versus $H_a : \delta > 0$ (i.e. $\delta_0 = 0$). The required sample size to reject H_0 at level $\alpha = 0.025$ and to detect a true difference $\delta = d_0 = 2$ given $\sigma = s_0 = 6$ is 143 patients per group at a 80% power and 191 patients per group at a 90% power. The predictive probabilities associated with these numbers of patients are only 66.6% and 72.1%. The numbers of patients per group to achieve the 80% and 90% predictive probabilities are 330 and 1457 respectively, thus a gain of only 10% in predictive probability results in dramatically increased costs. The maximum predictive probability that can be achieved with an infinite sample size is $1 - Pr(t_{\nu_0} > \frac{d_0}{s0\sqrt{\frac{2}{n_0}}}) = 95.1\%$

Brown *et al.* [2]have suggested calculating the predictive probability of achieving a positive outcome, by considering either the clinically relevant values of the true treatment effect ($\delta \geq \Delta_{min}$) or the possible values under the alternative hypothesis ($\delta > 0$). This approach is

even more demanding. However, the idea of restricting the possible values of the parameters to only the relevant ones is inappropriate, because, as the true treatment effect is unknown, nobody would be able to distinguish the successful trials, where the true treatment is relevant from those where it is not.

Finally, it should be emphasized that there still remains great uncertainty as to the appropriate sample size since there is a great uncertainty about the true power. Spiegelhalter *et al.* [6] suggested deriving by Monte-Carlo simulation the predictive distribution of the usual power from the prior distributions of the parameters δ and σ^2. In our previous numerical example, using the same prior distributions based on the previous data and considering 143 patients per group for the planned trial, the median of the simulated predictive distribution of the power is very close to 80%. The 95% highest predictive probability interval of the simulated predictive distribution of the power is $[0.9\%; 100\%]$, which illustrates the huge uncertainty about the true power, and consequently about the appropriate sample size.

In conclusion, while the frequentist approach assumes fixed values of the parameters, which could be implausible, the predictive power method provides a more realistic evaluation of the chances of rejecting H_0. Therefore this approach might prevent overenthusiastic sponsors from deluding themselves about the real success of the trials. It also has the undeniable merit of ensuring that the sponsors pay careful attention as to how the prior information is elicited. However, even if this method is a valuable aid to planning the trial, it is rarely used to justify the required sample size in a regulatory context. Firstly, resources are often limited. Secondly, if a company is prepared to enrol a larger sample size, the criterion for justifying this sample size is unlikely to be the high predictive power of achieving a successful outcome, but rather the ability to detect an effect of lesser size with controlled power. These are the main reasons why predictive power-based methods are very rarely used in the planning of drug trials.

Interval length methods

These methods are based on another approach for determining the sample size, which is to control the length of the interval estimates of the parameter of interest.

In the Bayesian framework, different criteria have been proposed in the literature (see Adcock [10], for a detailed review). The most common are the average length [11-12], the average coverage [11-12] and the worst outcome criteria [11-13] and can be briefly summarised as follows.

For the average length criterion, the sample size n is determined so that the mean length of the $100(1-\alpha)\%$ highest posterior density intervals weighted by their predictive probabilities is at most L. For the average coverage criterion, the sample size n is chosen so that the mean coverage of the highest posterior density intervals of fixed length L, weighted by their predictive probabilities, is at least $1-\alpha$. In this latter approach, the interval length L is considered fixed, and thus the coverage is random for each sample, which is rather unconventional. As Joseph and Bélisle [12] have acknowledged, 'cautious investigators may not be satisfied with the average assurances provided' by the two previous criteria. They suggested a more conservative approach based on the worst outcome criterion: the sample size is found such that there is a high $\gamma\%$ assurance (say 90%, for example) that the length of the $100(1-\alpha)\%$ highest posterior density intervals is at most L. The high assurance is obtained by ensuring that the length of these intervals is at most L for all the samples defined by the $\gamma\%$ highest predictive density region.

However, whichever criterion is chosen, comments can be made about this interval length-based approach. Firstly, when calculating the necessary sample sizes according to these criteria, the posterior distribution $p(\delta|x)$ should depend only on the planned trial data and not on prior information to satisfy regulatory expectations. Secondly, the criterion, that is, the interval length, on its own, is insufficient to make a claim for drug approval. Usually the goal of a comparative drug trial is to consider simultaneously the precision and the location of the treatment effect estimate, except in some very particular contexts in drug development, where the precision of the treatment effect estimate is the primary aim. In the context of clinical trials, precision must be considered in connection with location and therefore with the hypotheses tested, and consequently also with the associated power. Finally, all these methods rely on the specification of the desired interval length L, and therefore on its clinical interpretation. The length L cannot be easily justified on a clinical basis, except if its interpretation is related to the predictive chance of success, that is, the power. Thus these methods address the problem of SSD in drug trials in an indirect and clumsy manner and seem to be of little interest in this particular context. It is worth noting however that there are specific situations, such as subgroup analyses, where the power cannot be controlled at a high level, in which case, the achieved precision in a specific subgroup analysis (cf. Grouin *et al.* [25]) could bring additional clarity to the justification of the planned sample size.

Decision-theoretic methods

Lindley [14-15] and Bernardo [16] advocated that the SSD should be addressed as a decision problem to guarantee coherence. The method proposed by Lindley relies on the maximization of the expected utility in the Bayesian framework. The aim of a Bayesian decision procedure is to choose, on the basis of some utility function, one of the possible actions $a \in A$ whose consequences depend on some unknown parameter, δ, i.e. the true treatment effect in the context of clinical trials.

The utility of performing a trial of size n, observing data x and choosing a if δ is true, is defined as a function of these four quantities $u(a, \delta, x, n)$. For example, Raiffa and Schlaifer [17] proposed working with a utility function which does not depend on x and is expressed as the difference between a gain function $g(a, \delta)$ and a cost function $c(n)$,

$$u(a, \delta, x, n) = g(a, \delta) - c(n)$$

The gain and cost functions are usually expressed in the same unit, e.g. a monetary one. The cost is often expressed as a linear function of n, $c(n) = c_0 + c_1 n$, where c_0, c_1 are the set-up costs of the trial and the cost per patient respectively. Various gain functions have been proposed (see [18-20], for example) according to the context, the most common is the '0-1' gain according to the values of δ.

Coherence implies that the optimal sample size is the value of n which maximizes the expected utility

$$u^*(n) = \int \left(max_{a \in A} \left\{ \int u(a, \delta) p(\delta|x) d\delta \right\} \right) p(x) dx - c(n)$$

With this approach, the regulator faces a lot of issues that currently remain unsolved. As underlined by Joseph [21], the relevance and appropriateness of the chosen utility have to be proven. The utility may not capture the important features of the problem in terms of gains

or costs. For example, suppose there are multiple risks and benefits, some of which may be unforeseen, it will then be difficult to set costs and gains.

Even if a reasonable and functionally simple utility is available, this will always depend on constants which have to be guessed or anticipated. A wrong guess will lead to a wrong sample size.

Above all, costs and benefits are typical sponsor concerns but not necessarily regulatory ones. If a sponsor is interested in developing a new drug, their financial interest would be to set up a trial and to find the optimal sample size to maximize the utility of this trial. However regulatory priorities are to obtain enough evidence on efficacy and safety endpoints for drug approval, and reaching an agreement on a particular utility would be very difficult. It is easier and sometimes more realistic for regulators to agree on the required degree of accuracy so that useful information on efficacy and safety endpoints can be provided to all involved in the process of reviewing.

Lindley [15] argued strongly against this viewpoint in his response to Joseph and Wolfson [21]: 'The fact that virtually all sample size calculations that are performed today are not based on the maximization of expected utility is due to users unfamiliarity with the procedure in comparison with the fixation on probabilities of error that has become so ingrained in statisticians thinking ... Utilities are difficult in practice because people have failed to put enough research effort into their evaluation. A practical advantage of utilities is that they force us to compare things that we are frequently reluctant to consider together. In society we usually fail to appreciate the cost of keeping people alive, or to compare the costs of one medical treatment with another'.

The decision-based method is acknowledged to be a valuable tool for the sponsor to determine an optimal sample size to maximize the return on their investment, and it may help them to allocate limited clinical resources to different clinical projects on an informed basis. Although Lindley's concerns are important, they shouldn't dictate, for a drug approval, the size of a clinical trial which should be determined on a clinical basis only. If the decision-theoretic method leads to a required sample size larger than that based on the frequentist power, the regulators will be unlikely to object to this larger sample size. However, should the sample size turn out to be smaller, the regulators may not be convinced by the results based on such a sample size.

5. DISCUSSION

To summarize, the Bayesian methods for sample size determination, in the context of a non-sequential phase III trial, can help the sponsor in three ways: by assessing the realistic chances of getting a successful outcome, by eliciting prior information and by enabling efficient allocation of resources.

However, from a regulatory viewpoint, the value of some Bayesian methods, especially interval-based and decision-based methods, is highly debatable for all the reasons given in Section 4. In contrast, the value of the predictive power method is much more obvious in the context of the regulatory submission, and is discussed as follows.

It is worth remembering that, when the SSD is based on the minimal relevant effect, it is always acceptable to the regulators. However, most sponsors still base their calculations on the anticipated effect, taking advantage of information based on previous data, and not on

the minimal effect which requires a larger sample size. The frequentist power method could be misleading in assessing the chances of getting a positive outcome because it may be relying on an implausible value of the true effect. In contrast the predictive power method, which relies on the prior distribution of all the possible values of the true effect, has the advantage of making the sponsor aware of the realistic chances of a successful trial.

Therefore, from both regulatory and ethical viewpoints, when the SSD is based on an anticipated effect, it would be logical to choose the sample size to satisfy the predictive probability requirement rather than the simple power one. However, as the predictive power method requires larger sample sizes than those based on the frequentist power, predictive probabilities are very rarely/never used by sponsors when planning and justifying the sample sizes of the trials. In practice this means that if the regulatory bodies or ethical committees made the predictive probability requirement compulsory, the necessary sample sizes required for drug approval would increase significantly, thus increasing the cost of obtaining enough conclusive evidence for drug approval, possibly threatening future drug development. If the regulatory bodies and ethical committees agreed to acknowledge that the frequentist power method is an inadequate tool for a realistic evaluation of the chances of achieving the primary objective of the trial, then predictive probabilities lower than the conventional 80% threshold could be accepted in order to permit more reasonable sample sizes, i.e. similar to those required by the frequentist power method.

It is worth noting that the maximum achievable predictive probability of a successful outcome, i.e. one obtained with an infinite sample size, depends on the precision of the anticipated effect based on previous data. The less precise the previous data are, the lower the maximum predictive probability is. This could be considered a serious restriction to the use of predictive probabilities as a planning tool, especially as the population and design characteristics of the previous trial could have differed substantially from those of the planned trial. However, it should be remembered that the relationship between predictive probabilities and sample sizes increases rapidly before reaching a plateau, and thus a reasonable sample size can be selected for a predictive probability much lower than the achievable maximum.

Finally, the issue of sample size determination has only been addressed throughout this paper in the context of a fixed non-sequential trial. As the anticipated effect is based on previous data derived from trials where the experimental conditions could have been different from those of the planned trial, it may not be a sound basis for an accurate guess at the true effect. Thus, the need for a better educated guess has encouraged recent research into adaptive designs (cf. Jennison and Turnbull [26], for example). For example, designs which allow internal modifications of the sample size in the mid-course of the trial, given the estimated treatment effect and the nuisance parameters, these sample size modifications are more efficient because they depend on more reliable and relevant data. Thus, these specific designs appear to be very promising, and either approach, Bayesian or frequentist, could be used beneficially to reassess the sample size. However, although recently developed frequentist methods are intended to control type I and type II errors, it remains unclear whether these designs may induce consciously or unconsciously operational biases in the monitoring of the trial. Furthermore the efficiency aimed at by these methods, in terms of sample size decisions, may be illusory given the fact that the safety evaluation needs a larger amount of data. Therefore there is a clear need for more published data and discussion by regulators and sponsors to assess the validity of such designs.

For these reasons, sample size determination in a fixed non-sequential trial remains a relevant

issue. In this specific context, the Bayesian methods that allow for the uncertainty of the parameters seem to address in a natural way the problem of how to improve the guess at the true treatment effect.

ACKNOWLEDGEMENT

REFERENCES

1. ICH E9 Expert Working Group. Statistical principles for clinical trials: ICH harmonised tripartite guideline. *Statistics in Medicine* 1999; **18**: 1905-1942.
2. Brown BW, Herson J, Atkinson EN, Rozell ME. Projection from previous studies - a Bayesian and frequentist compromise. *Controlled Clinical Trials* 1987; **8**: 29-44.
3. Spiegelhalter DJ, Freedman LS. A predictive approach to selecting the size of a clinical trial, based on subjective clinical opinion. *Statistics in Medecine* 1986; **5**: 1-13.
4. Spiegelhalter DJ, Freedman LS, Parmar MK. Applying Bayesian ideas in drug development and clinical trials. *Statistics in Medecine* 1993; **12**: 1501-1517.
5. Spiegelhalter DJ, Freedman LS, Parmar MK. Bayesian approaches to randomized trials. *Journal of the Royal statistical society, series A* 1994; **157**: 357-416.
6. Spiegelhalter DJ, Abrams KR, Myles JP. Bayesian Approaches to clinical trials and health care evaluation. *New York: Wiley* 2004.
7. Lecoutre B. Two useful distributions for Bayesian predictive procedures under normal models. *Journal of Statistical Planning and Inference* 1999; **77**: 93-105.
8. Lecoutre B. Bayesian predictive procedures for designing and monitoring experiments. *In Bayesian Methods with Applications to Science, Policy and Official Statistics, Luxembourg: Office for Official Publications of the European Communities* 2001; 301-310.
9. Grouin J-M, Lecoutre B. Probabilités prédictives: Un outil pour la planification des expériences. *Revue de Statistique Appliquée* 1996; **44 (1)** : 21-35.
10. Adcock CJ. Sample size determination : a review. *The Statistician* 1997; **46 (2)**: 261-283.
11. Joseph L, Wolfson DB, Du Berger R. Some comments on Bayesian sample size determination. *The Statistician* 1995; **44**: 167-171.
12. Joseph L, Bélisle P. Bayesian sample size determination for normal means and differences between normal means. *The Statistician* 1997; **46 (2)**: 209-226.
13. Pham-Gia T, Turkkan N. Sample size determination in Bayesian Analysis. *The Statistician* 1992; **41**: 389-397.
14. Lindley DV. The choice of sample size. *The Statistician* 1997; 46 (2): 129-138.
15. Lindley DV. The choice of sample size - a reply to the discussion. *The Statistician* 1997; **46 (2)**: 163-166.
16. Bernardo J. Statistical inference as a decision problem: the choice of sample size. *The Statistician* 1997; **46 (2)**: 151-153.
17. Raiffa H, Schlaifer R. Applied statistical decision theory. *Boston: Harvard University Graduate School of Business Administration* 1961.
18. Claxton K, Posnett J. An economic approach to clinical trial design and research priority-setting. *Health Economics* 1996; **5**: 513-524.
19. Hornberger J. Introduction to Bayesian reasonning. *International Journal of Technology Assessment in Health care* 2000; **17**: 9-16.
20. Gittins J, Pezeshk H. How large should a clinical trial be? *The Statistician* 2000; **49**: 177-187.
21. Joseph L, Wolfson DB. Interval-based versus decision theoretic criteria for the choice of sample size. *The Statistician* 1997; **46 (2)**: 145-149.
22. Whitehead J. 'The case for frequentism in clinical trials. *Statistics in Medecine* 1993; **12**: 1405-1413.
23. Press SJ. Subjective and objective Bayesian statistics. *Wiley* 2003.
24. Berger J. The case for objective Bayesian analysis. *Bayesian Analysis* 2004; **1**, 1: 1-17.
25. Grouin J-M, Coste M, Lewis J. Subgroup analyses in randomised clinical trials: statistical and regulatory issues. *Journal of Biopharmaceutical Statistics* 2005; **15**, 5: 869-882.
26. Jennison C, Turnbull BW. Mid-course sample size modification in clinical trials based on the observed treatment effect. *Statistics in Medicine* 2003; **22**: 971-993.